Auditory Perceptual Organization

Synonyms

Auditory scene analysis, auditory perceptual grouping, sound source separation, cocktail party problem

Definition

The process of extracting acoustic features from sound waves and partitioning them into meaningful groups

Detailed Description

Introduction

Traveling pressure waves (ie. sounds) are produced by the movements or actions of objects. So sounds primarily convey information about *what is happening* in the environment. In addition, some information about the structure of the environment and the surface features of objects can be extracted by determining how the original (self-generated or exogenous) sounds are filtered or distorted by the environment (e.g. the notion of "acoustic daylight," (Fay 2009)). In this article we consider how the auditory systems processes sound signals to extract information about the environment and the objects within it.

The auditory system faces a number of specific challenges which need to be considered in any account of perceptual organization. 1) Sounds unfold in time; we can't (normally) go back to re-examine them. Therefore information must be extracted and perceptual decisions made in a timely manner. 2) The information contained within sounds generally requires processing over many timescales in order to extract their meaning (Nelken 2008). For example, a brief impulsive sound may tell the listener that two objects have been in collision, but a series of such sounds is needed in order for the listener to know that someone is clapping rather than walking. 3) Many objects of interest generate sounds intermittently. Therefore some means for associating temporally discontiguous events is required. 4) Sound pressure waves are additive; what the ear receives is a combination of all concurrently active sound sources and their reflections off any hard surfaces. Many animals and birds communicate acoustically in large social groups, making the problem of source separation particularly tricky (Bee 2012). Despite these challenges, if the auditory system is to provide meaningful information about individual objects in the environment (e.g. potential mates or aggressors) it needs to partition the acoustic features into meaningful groups, a process known as *auditory perceptual organization* or *auditory scene analysis* (Bregman 1990).

Grouping Principles

Auditory Events

Natural environments typically contain many concurrent sound sources, and even isolated sounds can be rather complex; e.g. animal vocalizations contain many different frequency components, and both the frequencies of the components and their amplitudes can vary within a single sound. The problem for the auditory system is to find some way of correctly associating the features which originate from the same sound source. The classical view of this process is that cochlea decomposes the incoming composite sound waveform into its spectral components, generating a topographically organised array of signals which sets up the *cochleotopic* (or *tonotopic*) organization found throughout most of the auditory system, up to and including the primary auditory cortex (Zwicker and Fastl 1999). Other low-level features such as onsets, amplitude and frequency modulations, and binaural differences are extracted sub-cortically and largely independently within each frequency channel (Oertel, Fay et al. 2002). These acoustic features are bound together to form auditory *events* (Bertrand and Tallon-Baudry 2000, Zhuo and Yu 2011) or *tokens* (Shamma, Elhilali et al. 2011), i.e. discrete sounds that are localised in time and perceived as originating from a single sound source (Ciocca 2008). Events are subsequently grouped sequentially into patterns, streams, or perceptual objects.

Gestalt Grouping Principles

Perceptual decisions regarding the *causes* of the signals received by the sensors must in general be made with incomplete information (Brunswik 1955). Therefore, potential solutions need to be constrained in some way; e.g. by knowledge about likely sound sources (Bar 2007), or by expectations arising from the recent context (Winkler, Denham et al. 2012). In his seminal book, Bregman (1990) pointed out that many such constraints had already been identified by the Gestalt school of psychology (Köhler 1947) early in the 20th century. The core observation of Gestalt psychology was that individual features form larger perceptual units, which have properties not present in the separate components (von Ehrenfels 1890); and conversely, that the perception of the components is influenced by the overall perceptual structure (Wertheimer 1912). Focussing primarily on visual stimuli the Gestalt psychologists described the following grouping principles (laws of perception); here discussed in terms of auditory grouping.

a) **Good continuation**: smooth continuous changes in perceptual attributes favour grouping, while abrupt discontinuities are perceived as the start of something new. This principle can operate both within and between individual events.
b) **Similarity**: similarity between the perceptual attributes of successive events (e.g. pitch, timbre, location) promotes grouping (Bregman 1990, Moore and Gockel 2002, Moore and Gockel 2012). Similar to the perception of visual motion (Weiss, Simoncelli et al. 2002), it appears that it is not so much the raw difference that is important, but rather the *rate of change*; the slower the rate of change between successive sounds the more similar they are judged (Winkler, Denham et al. 2012). In other words in the auditory modality *Similarity* and *Good continuation* may be equivalent.
c) **Common fate**: correlated changes in features promote grouping; recently formalised as *temporal coherence*; i.e. feature correlations within time windows that span periods longer than individual events (Elhilali, Ma et al. 2009, Shamma, Elhilali et al. 2011).
d) **Disjoint allocation (or belongingness)**: refers to the principle that each element of the sensory input is only assigned to one perceptual object; e.g. exclusive border assignment in Rubin's face-vase illusion. However, although generally true, this principle is sometimes violated in auditory perception; e.g. in *duplex perception*, the same sound component can contribute to the perception of a complex sound as well as being heard separately (Rand 1974, Fowler and Rosenblum 1990).
e) **Closure**: objects tend to be perceived as whole even if they are not complete; e.g. a glide (a tone with a rising or falling pitch) is perceived as continuing through a masking noise if its offset is masked by the noise (Miller and Licklider 1950, Riecke, Van Opstal et al. 2008). This applies more generally to the perception of global patterns (or 'Gestalts'); e.g. individual notes are subsumed into a melodic pattern (McDermott and Oxenham 2008), predictable individual speech sounds are perceived as present even if they are masked by a noise or missing (Warren, Wrightson et al. 1988). The auditory system is extraordinarily sensitive to repeating patterns and appears to readily use this cue to parse complex scenes (Winkler 2007, McDermott, Wrobleski et al. 2011).

An important concept that emerges from the idea of a 'Gestalt' as a pattern is that of predictability. In the case of auditory perception this refers to expectancies about sound events that have not yet occurred. By detecting patterns (or feature regularities) in the acoustic input the brain can construct representations that allow it to anticipate or 'explain away' (Pearl 1988) future events. In this way Gestalt theory connects to the ideas of unconscious inference (Helmholtz 1885), and perception as hypothesis formation (Gregory 1980).

Auditory Objects

While visual objects are widely accepted as fundamental representational units, the notion of an auditory object is less well established and there is as yet no universal agreement on how an auditory object should be defined;

e.g. see (Kubovy and Van Valkenburg 2001, Griffiths and Warren 2004, Winkler, van Zuijen et al. 2006, Shinn-Cunningham 2008). Based on the Gestalt principles and ideas of perceptual inference, outlined above, Winkler, Denham et al. (2009) proposed a definition of an auditory perceptual object as a predictive representation, constructed from feature regularities extracted from the incoming sounds. These object representations are temporally persistent and encode distributions over featural and temporal patterns, determined by the current context. The consolidated object representation therefore refers to *patterns of sound events*; individual sound events are processed within the context of the whole to which they belong. This definition of an auditory perceptual object is compatible with the definition of an auditory stream, as a coherent sequence of sounds separable from other concurrent or intermittent sounds (Bregman 1990). However, whereas the term 'auditory stream' refers to a phenomenological unit of sound organization, with separability as its primary property, the definition proposed by Winkler, Denham et al. (2009) emphasizes the extraction and representation of the unit as a pattern with predictable components (Winkler, Denham et al. 2012). While the usage of the term object is not universally accepted within the auditory domain, we will use it in this article as defined by Winkler, Denham et al. (2009).

Auditory Scene Analysis

In order to determine the perceptual qualities of individual sound events the brain must first bind their component features even though the number of concurrent auditory objects and which features belong to each is unknown *a priori*; this must be inferred incrementally from the on-going sensory input. Therefore, it is clear that the auditory system needs to use (top-down) contextual information to guide its grouping decisions, and some means for evaluating these decisions and revising them in the event that they prove to be incorrect. In the currently most widely accepted framework describing perceptual sound organization, Auditory Scene Analysis, Bregman (1990) proposes two separable processing stages. The first stage is suggested to be concerned with partitioning sound events into potential groups based primarily on featural similarities and differences. The second stage, within which prior knowledge and task demands exert their influence, is a competitive process between candidate organizations that determines which one is perceived. Within this framework there are two types of grouping: *simultaneous* grouping based on concurrent cues, and *sequential* grouping based on contextual temporal cues. For the reasons outlined above, these two are not really distinct (simultaneous cues are influenced by prior sequential grouping, e.g. (Darwin, Hukin et al. 1995, Bendixen, Jones et al. 2010), just as sequential grouping is influenced by the perceptual qualities of individual events (simultaneous grouping) (Bregman 1990); nevertheless, they provide a useful starting point for models of auditory scene analysis.

Simultaneous Grouping

In the absence of sequential grouping cues there are some features which automatically trigger the formation of individual sound events; for reviews see (Darwin and Carlyon 1995, Ciocca 2008). Common onsets and offsets form clear temporal boundaries, and the strategy adopted by the auditory system is to match onsets to offsets (including similarities between features and temporal proximity) in order to segregate perceptual events (Nakajima, Sasaki et al. 2000). Harmonicity (i.e. the presence of frequency components which are integer multiples of a common fundamental frequency) is another important grouping cue (Darwin and Carlyon 1995). For example, when one component of a complex harmonic tone is mistuned listeners perceive two concurrent sounds, a complex tone consisting of the harmonically related components and a pure tone, corresponding to the mistuned component (Moore, Glasberg et al. 1986). However, not all acoustic features trigger concurrent grouping; e.g. even if a subset of the frequency components of a single sound event are linked by a common location cue (common interaural time difference) they will not be perceptually segregated from the other components on the basis of this cue alone (Culling and Summerfield 1995).

Another important strategy for segregating sound events is template matching. If people have prior knowledge of events, then it is possible to hear them out. This effect was exploited in the many double vowel experiments (i.e. stimuli composed of a combination of two vowel sounds) used to test the influence of different acoustic features, e.g. (Assmann and Summerfield 1990, Summerfield and Assmann 1991), and even in the absence of featural differences, it was shown that known vowel sounds can be identified well above chance (Assmann and Summerfield 1989). This template matching phenomenon appears to be rather general and applies to any sound that is repeated. The auditory system is very sensitive to repetition (Teki, Chait et al. 2011). If a previously unheard sound is repeated against a different background, then it can be segregated and identified significantly above chance, even with only a single repetition, and even if many of usual grouping cues are absent (McDermott, Wrobleski et al. 2011). Similarly, arbitrary repeated noise segments can be rapidly learnt within a few trials (Agus, Thorpe et al. 2010).

Models of Event Formation

Many models have been developed to investigate simultaneous grouping and the segregation of perceptual events, e.g. see models described in Wang and Brown (2006). A model of auditory saliency which used low level cues of spectral and temporal contrast to detect salient events in continuous noisy soundscapes predicted human event detection very well (Kayser, Petkov et al. 2005). Temporal contrasts effectively highlight onsets and offsets, while spectral peaks carry information about the resonances of sound sources and to some extent their identity (von Kriegstein, Smith et al. 2007). The segregation of overlapping events using pitch cues has been widely explored (c.f. Pitch Perception, Models), e.g. segregation of two concurrent vowel sounds is enhanced if they have different pitches (de Cheveigne, McAdams et al. 1995). The segregation of events using repetition was shown to be possible in principle by using a combination of cross-correlation and averaging to incrementally build a representation of the repeated target (McDermott, Wrobleski et al. 2011). Because of the importance of longer term context on grouping, none of these models provide general solutions to the problem of auditory scene analysis; nevertheless, they provide important building blocks in this process.

Sequential Grouping

Sequential grouping generally conforms to the Gestalt principles of *similarity/good continuation* and *common fate.* In contrast to concurrent grouping, sequential grouping is necessarily based on some representation of the preceding sounds; for reviews, see (Moore and Gockel 2002, Carlyon 2004, Haykin and Chen 2005, Snyder and Alain 2007, Ciocca 2008, Shamma and Micheyl 2010, Shamma, Elhilali et al. 2011, Moore and Gockel 2012). Most studies of this class of grouping have used sequences of discrete sound events to investigate the influences of acoustic features and temporal structure. In the most widely used experimental approach (termed the auditory streaming paradigm), sequences of alternating sound events differing in some feature(s) are presented to listeners (van Noorden 1975). When the feature separation is small and/or they are delivered at a slow pace, listeners predominantly hear a single *integrated* stream containing all the sounds. With large feature separation and/or fast presentation rates, listeners report hearing the sequence separate out into two *segregated* streams. In this there is a cue trade-off: smaller feature differences can be compensated with higher presentation rates, and *vice versa* (van Noorden 1975). Differences in various auditory features, including frequency, pitch, loudness, location, timbre, and amplitude modulation, have been shown to support auditory stream segregation (Vliegen and Oxenham 1999, Grimault, Bacon et al. 2002, Roberts, Glasberg et al. 2002). Thus it appears that sequential grouping is based on perceptual similarity, rather than on specific low-level auditory features (Moore and Gockel 2002, Moore and Gockel 2012). Temporal structure has also been suggested as a key factor in segregating streams either by guiding attentive grouping processes (Jones 1976, Jones, Kidd et al. 1981, Large and Jones 1999) or through temporal coherence that binds correlated component features in the auditory input (Elhilali, Ma et al. 2009, Shamma and Micheyl 2010, Shamma, Elhilali et al. 2011, Shamma, Elhilali et al. 2013).

Models of Auditory Streaming

Early models of auditory streaming, e.g. (Beauvois and Meddis 1991), focussed on the relationship between frequency differences and event rate and the proposal that streaming could be explained almost exclusively by peripheral channelling mechanisms (Hartmann and Johnson 1991) or the degree of overlap between neural responses to each of the alternating tones, e.g. (McCabe and Denham 1997). In these models the perceptual decision was represented by levels of activation across a spatial array of neurons; see also (Micheyl, Tian et al. 2005) for a similar interpretation of neural activity in primary auditory cortex. A different approach in which grouping is signalled by temporal correlations within network responses was proposed by Wang, Brown and colleagues (Brown and Wang 2006, Wang and Chang 2008). For example, the model proposed by Wang and Chang (2008) consists of a 2-dimensional array of oscillators with one dimension representing frequency and the other external time. Units are connected by local excitatory connections and by global inhibition. Characteristic results of classical auditory streaming experiments (van Noorden 1975) are simulated by including strong local excitatory connections (encouraging synchronisation) and weaker long range connections (which are easily overcome by inhibition and therefore encourage desynchronisation). Sensitivity to event rate is modelled by dynamic weight adjustments. However, while the representation of grouping is different from the models previously outlined, this model also depends on peripheral channelling and the degree of overlap in the incoming activity patterns to determine its grouping decision.

A similar focus on temporal coherence (in this case the average correlation within a sliding window 50-500ms in duration) is seen in the model of streaming proposed by Elhilali and colleagues, e.g. (Elhilali and Shamma 2008, Shamma, Elhilali et al. 2011)[1]. The computational model developed by Elhilali and Shamma (2008) extracts multiple features from the incoming acoustic input including frequency, pitch, direction and spectral shape and assigns the resulting activity patterns to one of two clusters which come to represent the properties of the events in each stream. The temporal coherence measure is used to determine which components should be grouped. The clusters compete to incorporate each event, and the winning cluster uses the event features (as determined by the grouping process) to refine its representation. These correlation-based models overcome a problem faced by the population separation account of streaming (Micheyl, Tian et al. 2005) that predicted widely separated components would be segregated even if they overlapped in time, which is not the case (Elhilali, Ma et al. 2009). They also provide a means for binding the component features of an event, not considered in the earlier models. Later refinements to the temporal coherence account of streaming (Shamma, Elhilali et al. 2011, Shamma, Elhilali et al. 2013), included the strong claims that a) feature binding occurs only with attention; i.e. attention is responsible for grouping features that belong to the foreground object, cf. (Treisman 1998), and b) all other features remain ungrouped in an undifferentiated background. However, the proposed role of attention in feature binding has long been debated in the visual domain, e.g. (Duncan and Humphreys 1989), and it is not consistent with the results of experiments testing feature binding in the absence of attention by recording auditory event-related potentials (AERP) in response to rare feature combinations (Takegata, Brattico et al. 2005, Winkler, Czigler et al. 2005).

Competition and Selection

The models described above all conform to the assumptions that in response to alternating two tone sequences a) auditory perception always starts from the integrated organization, and b) that eventually a stable final perceptual decision is reached (Bregman 1990). However, it has been found when listeners report their percepts continuously while listening to such sequences for long periods, that perception fluctuates between different perceptual organisations (Winkler, Takegata et al. 2005, Pressnitzer and Hupe 2006). Perceptual switching occurs in all listeners and for all combinations of stimulus parameters tested (Anstis and Saida 1985, Roberts, Glasberg et al. 2002, Denham and Winkler 2006, Pressnitzer and Hupe 2006, Schadwinkel and Gutschalk 2011, Denham, Gymesi et al. 2012), even combinations very far from the ambiguous region identified by van Noorden (1975). Furthermore, for stimuli with parameters that strongly promote segregation, participants often report hearing segregation *first* (Deike, Heil et al. 2012, Denham, Gymesi et al. 2012). It has also been found that perceptual organisations other than the classic integrated and segregated categories may be reported (Bendixen, Denham et al. 2010, Bendixen, Bőhm et al. 2012, Bőhm, Shestopalova et al. 2012, Denham, Gymesi et al. 2012, Szalárdy, Bendixen et al. 2012), showing that auditory perceptual organisation in response to alternating two tone sequences is multi-stable (Schwartz, Grimault et al. 2012).

[1] Note, figures 6 and 9 in this paper have incorrect colour scale labels (0% and 100%, interchanged); Shamma, S. A. and M. Elhilali (2013).

The notion of perceptual multistability is challenged by everyday subjective experience of a world perceived as stable and continuous, and by experimental results obtained by averaging over the reports of different listeners, which generally show that within the initial 5-15 s of two-tone sequence, the probability of reporting segregation monotonically increases (termed the build-up of auditory streaming) (but see Deike, Heil et al. (2012)). For these reasons it has been suggested that perceptual multistability observed in the auditory streaming paradigm may be simply a consequence of the artificial stimulation protocol used. However, there is a growing body of experimental data supporting the existence of multistability, and just as visual multistability has provided new insights into visual processing, e.g. (Kovacs, Papathomas et al. 1996), it seems likely that understanding spontaneous changes in the perception of unchanging sound sequences will help throw new light on auditory perception.

Modelling Multistability in Auditory Streaming

Multistability of auditory perceptual organisation cannot be explained by any of the theories or models outlined above, which all have essentially one fixed attractor. Models of visual multistability have a longer history, e.g. (Laing and Chow 2002, Shpiro, Moreno-Bote et al. 2009, van Ee 2009). These models typically contain three essential components (Leopold and Logothetis 1999): a) mutual inhibition between competing stimuli to ensure *exclusivity* (ie. perceptual awareness generally switches between the different alternatives rather than fusing them), b) adaptation to ensure the observed *inevitability* of perceptual switching (the dominant percept cannot remain dominant forever), and c) noise to account for the observed *stochasticity* of perceptual switching (successive phase durations are largely uncorrelated, and the distribution of phase durations resembles a gamma or log-normal distribution) (Levelt 1968). The questions for auditory multistability are what are the competing entities, and what form does this competition take in order to explain dynamic nature of perceptual awareness reported by listeners.

The computational model of auditory multistability proposed by Mill, Bőhm et al. (2013) is based on the idea that auditory perceptual organisation rests on the discovery of recurring patterns embedded within the stimulus, constructed by forming associations (links) between incoming sound events and recognising when a previously discovered sequence recurs and can thus be used to predict future events. These predictive representations, or *proto-objects* (Rensink 2000, Winkler, Denham et al. 2012), compete for dominance with any other proto-objects which predict the same event (a form of local competion), and are the candidate set of representations that have the potential to become the perceptual objects of conscious awareness. This model accounts for the emergence of, and switching between, alternative organisations; the influence of stimulus parameters on perceptual dominance, switching rate and perceptual phase durations; and the build-up of auditory streaming. In a new sound scene, the proto-object that is easiest to discover determines the initial percept. Since the time needed for discovering a proto-object depends largely on the stimulus parameters (i.e., to what extent successive sound events satisfy/violate the *similarity/good continuation* principle), the first percept strongly depends on stimulus parameters. However, the duration of the first perceptual phase is independent of the percept (Hupe and Pressnitzer 2012), since it depends on how long it takes for *other* proto-objects to be discovered (Winkler, Denham et al. 2012). The model also accounts for the different influences of similarity and closure on perception; the rate of perceptual change (similarity/good continuation) determines how easy it is to form the links between the events that make up a proto-object, while predictability (closure) does not affect the discovery of proto-objects, but can increase the competitiveness (salience) of a proto-object once it has been discovered (Bendixen, Denham et al. 2010).

Neural Correlates of Perceptual Organization

Neural responses to individual sounds are profoundly influenced by the context in which they appear (Bar-Yosef, Rotman et al. 2002). The question is to what extent the contextual influences on neural responses reflect the current state of perceptual organization. This question has been addressed by a number of studies ranging in focus from the single neuron level (c.f. stimulus specific adaptation) to large scale brain responses (c.f. auditory evoked potentials), and the results provide important clues about the processing strategies adopted by the auditory system.

Studies investigating single neuron responses to alternating tone sequences, e.g. (Fishman, Arezzo et al. 2004, Bee and Klump 2005, Micheyl, Tian et al. 2005, Micheyl, Carlyon et al. 2007), have shown an effect called *differential suppression*; i.e. at the start of the sequence the neuron responds to both tones but with time the response to one of the tones (typically corresponding to the best frequency of the cell) remains relatively strong, while the response to the other tone diminishes. Since neuronal sensitivity to frequency difference and presentation rate was found to be consistent with the classical van Noorden (1975) parameter space, it was claimed that differential suppression was a neural correlate of perceptual segregation (Fishman, Arezzo et al. 2004). This was supported by the finding that spike counts from neurons in primary auditory cortex predict an initial integration/segregation decision closely matching human perception (Micheyl, Tian et al. 2005, Bee, Micheyl et al. 2010). However, differential suppression does not account for perceptual multistability nor for the perception of overlapping tone sequences (Elhilali, Ma et al. 2009); therefore, while differential suppression may be necessary component of the auditory streaming process, it does not provide a complete explanation.

Auditory event-related brain potentials (AERPs) represent the synchronized activity of large neuronal populations, time-locked to some auditory event. Because they can be recorded noninvasively from the human scalp, they have been widely used to study the brain responses accompanying auditory stream segregation; c.f. Auditory event-related potentials, especially Long latency AERP responses. Three AERP components are of particularly relevance in this regard: a) the 'object-related negativity' (ORN) which signals the automatic segregation of concurrent auditory objects (Alain, Schuler et al. 2002), b) the amplitude of the auditory P1 and N1 which vary depending on whether the same sounds are perceived as part of an integrated or segregated organization (Gutschalk, Micheyl et al. 2005, Szalárdy, Bőhm et al. 2013), and c) the Mismatch Negativity (MMN; (Näätänen, Gaillard et al. 1978)) which has been used as an indirect index of auditory stream segregation, e.g. (Sussman, Ritter et al. 1999, Nager, Teder-Sälejärvi et al. 2003, Winkler, Sussman et al. 2003, Gutschalk, Micheyl et al. 2005).

The detection and representation of regularities by the brain, as indexed by the MMN, provided the basis for the definition of an auditory object proposed by Winkler, Denham et al. (2009). Using evidence from a series of MMN studies, they defined an auditory object as a perceptual representation of a possible sound source, derived from regularities in the sensory input (Winkler 2007, Winkler 2010) that has temporal persistence (Winkler and Cowan 2005) and can link events separated in time (Näätänen and Winkler 1999). This representation forms a separable unit (Winkler, van Zuijen et al. 2006) that generalises across natural variations in the sounds (Winkler, Teder-Salejarvi et al. 2003) and generates expectations of parts of the object not yet available (Bendixen, Schröger et al. 2009).

It should be pointed out that while traditional psychological accounts of auditory perceptual organisation implicitly or explicitly refer to representations of objects, there are models of auditory perception which are not concerned with positing a representation directly corresponding to auditory objects. The hierarchical predictive coding model of perception, e.g. (Friston and Kiebel 2009), includes predictive memory representations, which are in many ways compatible with the notion of auditory object representations (Winkler and Czigler 2012), but no explicit connection with object representations is made. Shamma and colleagues' temporal coherence model of auditory stream segregation (Elhilali and Shamma 2008, Elhilali, Ma et al. 2009, Shamma, Elhilali et al. 2011, Shamma, Elhilali et al. 2013) provides another way to avoid the assumption that object representations are necessary for determining sound organization, instead it is proposed that objects are essentially whatever occupies the perceptual foreground and exist only insofar as they *do* occupy the foreground. In summary, there is currently little consensus on the role of auditory object representations in perceptual organization and the importance placed on object representations by the various models and theories differs markedly.

fMRI studies of auditory streaming have found neural correlates in a number of brain regions. In one of the earliest studies, Cusack (2005) failed to find differential activity in auditory cortex corresponding to perceptual organisation into one or two streams, but he did find such activity in the intra-parietal sulcus, an area associated with cross-modal processing and object numerosity. Shortly afterwards (Wilson, Melcher et al. 2007) showed that auditory cortical activity increased with increasing frequency difference and that as the frequency difference increased, the cortical response changed from being rather phasic (i.e. far stronger at the onset of the sequence) towards a more sustained response throughout the stimulus sequence. Taking a closer look at the dynamics of cortical activity associated with perceptual switching, Kondo and Kashino (2009) showed that both auditory cortex and thalamus are involved, with an increase in thalamic activity preceding that in cortex associated with a switch from the non-dominant to the dominant percept, and conversely, an increase in cortical activity preceding that in thalamus associated with a switch from the dominant to the non-dominant percept. They also found differential activation in posterior insular cortex and in the cerebellum. Interestingly, activations in the cerebellum and thalamus are negatively correlated in auditory streaming, with the left cerebellar activation level increasing with the rate of perceptual switching and thalamus (medial geniculate) decreasing (Kashino and Kondo 2012). Consistent with these findings, Schadwinkel and Gutschalk (2011), using a different stimulus paradigm which allowed them to influence the timing of perceptual switching, found transient auditory cortical activation associated with perceptual switching, and a further transient activation in inferior colliculus; although whether the inferior colliculus is responsible for triggering switching or simply reflects the transient switching activation in cortex is not clear. In summary, neural correlates of auditory streaming have been found in many areas within the auditory system and beyond, suggesting that creating and switching between alternative perceptual organisations involves a broadly distributed network within the brain.

Conclusions and Open Questions

The Gestalt principles and their application to auditory perception instantiated in Bregman's (1990) two-stage auditory scene analysis framework provided the initial basis for understanding auditory perceptual organization, and recent proposals have extended this framework in interesting ways. Nevertheless, there remain many unanswered questions and there have been few, if any, attempts to build neuro-computational models capable of dealing with the complexity of real auditory scenes in which grouping and categorization cues are not immediately available; however, see (Yildiz and Kiebel 2011). Feedback connections are pervasive within the auditory system, including all stages of the subcortical system, yet to our knowledge no models include such connections. Although fMRI results are useful for identifying regional involvement detailed understanding of the neural circuitry involved in auditory perceptual organisation is sketchy, and the neural representations of auditory objects and perceptual organisation are unknown. Even the role of primary auditory cortex remains something of a mystery; e.g. see (Nelken, Fishbach et al. 2003, Griffiths, Warren et al. 2004); perhaps studying the switching of perceptual awareness between different representations in awake behaving animals will help to elucidate the representations and processing strategies adopted by cortex.

Cross-References/Related Terms

Stimulus specific adaptation, Auditory event related potentials

References

Agus, T. R., S. J. Thorpe and D. Pressnitzer (2010). "Rapid formation of robust auditory memories: insights from noise." Neuron 66(4): 610-618.

Alain, C., B. M. Schuler and K. L. McDonald (2002). "Neural activity associated with distinguishing concurrent auditory objects." J Acoust Soc Am 111(2): 990-995.

Anstis, S. and S. Saida (1985). "Adaptation to auditory streaming of frequency-modulated tones." Journal of Experimental Psychology: Human Perception and Performance 11: 257-271.

Assmann, P. F. and Q. Summerfield (1989). "Modeling the perception of concurrent vowels: vowels with the same fundamental frequency." J Acoust Soc Am 85(1): 327-338.

Assmann, P. F. and Q. Summerfield (1990). "Modeling the perception of concurrent vowels: vowels with different fundamental frequencies." J Acoust Soc Am 88(2): 680-697.

Bar-Yosef, O., Y. Rotman and I. Nelken (2002). "Responses of neurons in cat primary auditory cortex to bird chirps: effects of temporal and spectral context." J Neurosci 22(19): 8619-8632.

Bar, M. (2007). "The proactive brain: using analogies and associations to generate predictions." Trends Cogn Sci 11(7): 280-289.

Beauvois, M. W. and R. Meddis (1991). "A computer model of auditory stream segregation." Q J Exp Psychol A 43(3): 517-541.

Bee, M. A. (2012). "Sound source perception in anuran amphibians." Curr Opin Neurobiol 22(2): 301-310.

Bee, M. A. and G. M. Klump (2005). "Auditory stream segregation in the songbird forebrain: effects of time intervals on responses to interleaved tone sequences." Brain Behav Evol 66(3): 197-214.

Bee, M. A., C. Micheyl, A. J. Oxenham and G. M. Klump (2010). "Neural adaptation to tone sequences in the songbird forebrain: patterns, determinants, and relation to the build-up of auditory streaming." J Comp Physiol A Neuroethol Sens Neural Behav Physiol 196(8): 543-557.

Bendixen, A., T. M. Bőhm, O. Szalárdy, R. Mill, S. L. Denham and I. Winkler (2012). "Different roles of similarity and predictability in auditory stream segregation." J. Learning & Perception in press.

Bendixen, A., S. L. Denham, K. Gyimesi and I. Winkler (2010). "Regular patterns stabilize auditory streams." J Acoust Soc Am 128(6): 3658-3666.

Bendixen, A., S. J. Jones, G. Klump and I. Winkler (2010). "Probability dependence and functional separation of the object-related and mismatch negativity event-related potential components." Neuroimage 50(1): 285-290.

Bendixen, A., E. Schröger and I. Winkler (2009). "I heard that coming: event-related potential evidence for stimulus-driven prediction in the auditory system." J Neurosci 29(26): 8447-8451.

Bertrand, O. and C. Tallon-Baudry (2000). "Oscillatory gamma activity in humans: a possible role for object representation." Int J Psychophysiol 38(3): 211-223.

Bőhm, T. M., L. Shestopalova, A. Bendixen, A. G. Andreou, J. Georgiou, G. Garreau, P. Pouliquen, A. Cassidy, S. L. Denham and I. Winkler (2012). "Spatial location of sound sources biases auditory stream segregation but their motion does not." J. Learning & Perception in press

Bregman, A. S. (1990). Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MA, MIT Press.

Brown, G. J. and D. L. Wang (2006). Neural and Perceptual Modelling. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. D. L. Wang and G. J. Brown, Wiley/IEEE Press.

Brunswik, E. (1955). "Representative design and probabilistic theory in a functional psychology." Psychological Review 62(3): 193-217.

Carlyon, R. P. (2004). "How the brain separates sounds." Trends Cogn Sci 8(10): 465-471.

Ciocca, V. (2008). "The auditory organization of complex sounds." Front Biosci 13: 148-169.

Culling, J. F. and Q. Summerfield (1995). "Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay." J Acoust Soc Am 98(2 Pt 1): 785-797.

Cusack, R. (2005). "The intraparietal sulcus and perceptual organization." J Cogn Neurosci 17(4): 641-651.

Darwin, C. J. and R. P. Carlyon (1995). Auditory grouping. The handbook of perception and cognition: Hearing. B. C. J. Moore. London, Academic Press. 6: 387-424.

Darwin, C. J., R. W. Hukin and B. Y. al-Khatib (1995). "Grouping in pitch perception: evidence for sequential constraints." J Acoust Soc Am 98(2 Pt 1): 880-885.

de Cheveigne, A., S. McAdams, J. Laroche and M. Rosenberg (1995). "Identification of concurrent harmonic and inharmonic vowels: a test of the theory of harmonic cancellation and enhancement." J Acoust Soc Am 97(6): 3736-3748.

Deike, S., P. Heil, M. Böckmann-Barthel and A. Brechmann (2012). "The build-up of auditory stream segregation: a different perspective." Frontiers in Psychology 3.

Denham, S. L., K. Gymesi, G. Stefanics and I. Winkler (2012). "Multistability in auditory stream segregation: the role of stimulus features in perceptual organisation." Journal of Learning & Perception in press.

Denham, S. L. and I. Winkler (2006). "The role of predictive models in the formation of auditory streams." J Physiol Paris 100(1-3): 154-170.

Duncan, J. and G. Humphreys (1989). "Visual search and stimulus similarity." Psychological Review 96: 433=458.

Elhilali, M., L. Ma, C. Micheyl, A. J. Oxenham and S. A. Shamma (2009). "Temporal coherence in the perceptual organization and cortical representation of auditory scenes." Neuron 61(2): 317-329.

Elhilali, M. and S. A. Shamma (2008). "A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation." J Acoust Soc Am 124(6): 3751-3771.

Fay, R. (2009). "Soundscapes and the sense of hearing of fishes." Integr Zool 4(1): 26-32.

Fishman, Y. I., J. C. Arezzo and M. Steinschneider (2004). "Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration." J Acoust Soc Am 116(3): 1656-1670.

Fowler, C. A. and L. D. Rosenblum (1990). "Duplex perception: a comparison of monosyllables and slamming doors." J Exp Psychol Hum Percept Perform 16(4): 742-754.

Friston, K. and S. Kiebel (2009). "Predictive coding under the free-energy principle." Philos Trans R Soc Lond B Biol Sci 364(1521): 1211-1221.

Griffiths, T. D. and J. D. Warren (2004). "What is an auditory object?" Nat Rev Neurosci 5(11): 887-892.
Griffiths, T. D., J. D. Warren, S. K. Scott, I. Nelken and A. J. King (2004). "Cortical processing of complex sound: a way forward?" Trends Neurosci 27(4): 181-185.

Grimault, N., S. P. Bacon and C. Micheyl (2002). "Auditory stream segregation on the basis of amplitude-modulation rate." J Acoust Soc Am 111(3): 1340-1348.

Gutschalk, A., C. Micheyl, J. R. Melcher, A. Rupp, M. Scherg and A. J. Oxenham (2005). "Neuromagnetic correlates of streaming in human auditory cortex." J Neurosci 25(22): 5382-5388.

Hartmann, W. M. and D. Johnson (1991). "Stream Segregation and Peripheral Channeling." Music Perception: An interdisciplinary journal 9(2): 153-183.

Haykin, S. and Z. Chen (2005). "The cocktail party problem." Neural Comput 17(9): 1875-1902.
Helmholtz, H. v. (1885). On the sensations of tone as a physiological basis for the theory of music. London, Longmans, Green, and Co.

Hupe, J. M. and D. Pressnitzer (2012). "The initial phase of auditory and visual scene analysis." Philos Trans R Soc Lond B Biol Sci 367(1591): 942-953.

Jones, M. R. (1976). "Time, our lost dimension: Toward a new theory of perception, attention, and memory." Psychological Review 83: 323-355.

Jones, M. R., G. Kidd and R. Wetzel (1981). "Evidence for rhythmic attention." Journal of Experimental Psychology: Human Perception & Performance 7: 1059-1073.

Kashino, M. and H. M. Kondo (2012). "Functional brain networks underlying perceptual switching: auditory streaming and verbal transformations." Philos Trans R Soc Lond B Biol Sci 367(1591): 977-987.

Kayser, C., C. I. Petkov, M. Lippert and N. K. Logothetis (2005). "Mechanisms for allocating auditory attention: an auditory saliency map." Curr Biol 15(21): 1943-1947.

Köhler, W. (1947). Gestalt psychology: An introduction to new concepts in modern psychology. New York, Liveright Publishing Corporation.

Kondo, H. M. and M. Kashino (2009). "Involvement of the thalamocortical loop in the spontaneous switching of percepts in auditory streaming." J Neurosci 29(40): 12695-12701.

Kovacs, I., T. V. Papathomas, M. Yang and A. Feher (1996). "When the brain changes its mind: interocular grouping during binocular rivalry." Proc Natl Acad Sci U S A 93(26): 15508-15511.

Kubovy, M. and D. Van Valkenburg (2001). "Auditory and visual objects." Cognition 80(1-2): 97-126.

Laing, C. R. and C. C. Chow (2002). "A spiking neuron model for binocular rivalry." J Comput Neurosci 12(1): 39-53.

Large, E. W. and M. R. Jones (1999). "The dynamics of attending: How people track time-varying events." Psychological Review 106: 119–159.

Leopold, D. A. and N. K. Logothetis (1999). "Multistable phenomena: changing views in perception." Trends Cogn Sci 3(7): 254-264.

Levelt, W. J. M. (1968). On binocular rivalry. Paris, Mouton.

McCabe, S. L. and M. J. Denham (1997). "A model of auditory streaming." J. Acoust. Soc. Am. 101(3): 1611-1621.

McDermott, J. H. and A. J. Oxenham (2008). "Music perception, pitch, and the auditory system." Curr Opin Neurobiol 18(4): 452-463.

McDermott, J. H., D. Wrobleski and A. J. Oxenham (2011). "Recovering sound sources from embedded repetition." Proc Natl Acad Sci U S A 108(3): 1188-1193.

Micheyl, C., R. P. Carlyon, A. Gutschalk, J. R. Melcher, A. J. Oxenham, J. P. Rauschecker, B. Tian and E. Courtenay Wilson (2007). "The role of auditory cortex in the formation of auditory streams." Hear Res 229(1-2): 116-131.

Micheyl, C., B. Tian, R. P. Carlyon and J. P. Rauschecker (2005). "Perceptual organization of tone sequences in the auditory cortex of awake macaques." Neuron 48(1): 139-148.

Mill, R., T. Bőhm, A. Bendixen, I. Winkler and S. L. Denham (2013). "Competition and Cooperation between Fragmentary Event Predictors in a Model of Auditory Scene Analysis." PLoS Comput Biol in press.

Miller, G. A. and J. C. R. Licklider (1950). " The intelligibility of interrupted speech." Journal of Acousitcal Society of America 22: 167-173.

Moore, B. C., B. R. Glasberg and R. W. Peters (1986). "Thresholds for hearing mistuned partials as separate tones in harmonic complexes." J Acoust Soc Am 80(2): 479-483.

Moore, B. C. and H. E. Gockel (2012). "Properties of auditory stream formation." Philos Trans R Soc Lond B Biol Sci 367(1591): 919-931.

Moore, B. C. J. and H. E. Gockel (2002). "Factors influencing sequential stream segregation." Acta Acust. 88: 320-333.

Näätänen, R., A. W. K. Gaillard and S. Mäntysalo (1978). "Early selective attention effect on evoked potential reinterpreted." Acta Psychol 42: 313-329.

Näätänen, R. and I. Winkler (1999). "The concept of auditory stimulus representation in cognitive neuroscience." Psychol Bull 125(6): 826-859.

Nager, W., W. Teder-Sälejärvi, S. Kunze and T. F. Münte (2003). "Preattentive evaluation of multiple perceptual streams in human audition." Neuroreport 14(6): 871-874.

Nakajima, Y., T. Sasaki, K. Kanafuka, A. Miyamoto, G. Remijn and G. ten Hoopen (2000). "Illusory recouplings of onsets and terminations of glide tone components." Percept Psychophys 62(7): 1413-1425.

Nelken, I. (2008). "Processing of complex sounds in the auditory system." Curr Opin Neurobiol 18(4): 413-417.

Nelken, I., A. Fishbach, L. Las, N. Ulanovsky and D. Farkas (2003). "Primary auditory cortex of cats: feature detection or something else?" Biol Cybern 89(5): 397-406.

Oertel, D., R. R. Fay and A. N. Popper (2002). Integrative Functions in the Mammalian Auditory Pathway. New York, Springer-Verlag.

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, Morgan Kaufmann Publishers.

Pressnitzer, D. and J. M. Hupe (2006). "Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization." Curr Biol 16(13): 1351-1357.

Rand, T. C. (1974). "Letter: Dichotic release from masking for speech." J Acoust Soc Am 55(3): 678-680.
Rensink, R. A. (2000). "Seeing, sensing, and scrutinizing." Vision Res 40(10-12): 1469-1487.

Riecke, L., A. J. Van Opstal and E. Formisano (2008). "The auditory continuity illusion: a parametric investigation and filter model." Percept Psychophys 70(1): 1-12.

Roberts, B., B. R. Glasberg and B. C. Moore (2002). "Primitive stream segregation of tone sequences without differences in fundamental frequency or passband." J Acoust Soc Am 112(5 Pt 1): 2074-2085.

Schadwinkel, S. and A. Gutschalk (2011). "Transient bold activity locked to perceptual reversals of auditory streaming in human auditory cortex and inferior colliculus." J Neurophysiol 105(5): 1977-1983.

Schwartz, J. L., N. Grimault, J. M. Hupe, B. C. Moore and D. Pressnitzer (2012). "Multistability in perception: binding sensory modalities, an overview." Philos Trans R Soc Lond B Biol Sci 367(1591): 896-905.

Shamma, S., M. Elhilali, L. Ma, C. Micheyl, A. J. Oxenham, D. Pressnitzer, P. Yin and Y. Xu (2013). "Temporal coherence and the streaming of complex sounds." Adv Exp Med Biol 787: 535-543.

Shamma, S. A. and M. Elhilali (2013).

Shamma, S. A., M. Elhilali and C. Micheyl (2011). "Temporal coherence and attention in auditory scene analysis." Trends Neurosci 34(3): 114-123.

Shamma, S. A. and C. Micheyl (2010). "Behind the scenes of auditory perception." Curr Opin Neurobiol 20(3): 361-366.

Shinn-Cunningham, B. G. (2008). "Object-based auditory and visual attention." Trends Cogn Sci 12(5): 182-186.

Shpiro, A., R. Moreno-Bote, N. Rubin and J. Rinzel (2009). "Balance between noise and adaptation in competition models of perceptual bistability." J Comput Neurosci 27(1): 37-54.

Snyder, J. S. and C. Alain (2007). "Toward a neurophysiological theory of auditory stream segregation." Psychol Bull 133(5): 780-799.

Summerfield, Q. and P. F. Assmann (1991). "Perception of concurrent vowels: effects of harmonic misalignment and pitch-period asynchrony." J Acoust Soc Am 89(3): 1364-1377.

Sussman, E. S., W. Ritter and H. G. Vaughan, Jr. (1999). "An investigation of the auditory streaming effect using event-related brain potentials." Psychophysiology 36(1): 22-34.

Szalárdy, O., A. Bendixen, D. Tóth, S. L. Denham and I. Winkler (2012). "Modulation-frequency acts as a primary cue for auditory stream segregation." J. Learning & Perception in press.

Szalárdy, O., T. Bőhm, A. Bendixen and I. Winkler (2013). "Perceptual organization affects the processing of incoming sounds: An ERP study." Biological Psychology in press.

Takegata, R., E. Brattico, M. Tervaniemi, O. Varyagina, R. Naatanen and I. Winkler (2005). "Preattentive representation of feature conjunctions for concurrent spatially distributed auditory objects." Brain Res Cogn Brain Res 25(1): 169-179.

Teki, S., M. Chait, S. Kumar, K. von Kriegstein and T. D. Griffiths (2011). "Brain bases for auditory stimulus-driven figure-ground segregation." J Neurosci 31(1): 164-171.

Treisman, A. (1998). "Feature binding, attention and object perception." Philosophical Transactions of the Royal Society of London. Series B, Biological sciences 353: 1295-1306.

van Ee, R. (2009). "Stochastic variations in sensory awareness are driven by noisy neuronal adaptation: evidence from serial correlations in perceptual bistability." J Opt Soc Am A Opt Image Sci Vis 26(12): 2612-2622.

van Noorden, L. P. A. S. (1975). Temporal coherence in the perception of tone sequences. Doctoral dissertation, , Technical University Eindhoven.

Vliegen, J. and A. J. Oxenham (1999). "Sequential stream segregation in the absence of spectral cues." J Acoust Soc Am 105(1): 339-346.

von Ehrenfels, C. (1890). "Über Gestaltqualitäten (English "On the Qualities of Form")." Vierteljahrsschrift für wissenschaftliche Philosophie 14: 249-292.

von Kriegstein, K., D. R. Smith, R. D. Patterson, D. T. Ives and T. D. Griffiths (2007). "Neural representation of auditory size in the human voice and in sounds from other resonant sources." Curr Biol 17(13): 1123-1128.

Wang, D. L. and G. J. Brown (2006). Computational Auditory Scene Analysis: Principles, Algorithms, and Applications, Wiley/IEEE Press.

Wang, D. L. and P. S. Chang (2008). "An oscillatory correlation model of auditory streaming." Cognitive Neurodynamics 2: 7-19.

Warren, R. M., J. M. Wrightson and J. Puretz (1988). "Illusory continuity of tonal and infratonal periodic sounds." J Acoust Soc Am 84(4): 1338-1342.

Weiss, Y., E. P. Simoncelli and E. H. Adelson (2002). "Motion illusions as optimal percepts." Nat Neurosci 5(6): 598-604.

Wertheimer, M. (1912). "Experimentelle Studien über das Sehen von Bewegung." Zeitschrift für Psychologie 60.

Wilson, E. C., J. R. Melcher, C. Micheyl, A. Gutschalk and A. J. Oxenham (2007). "Cortical FMRI activation to sequences of tones alternating in frequency: relationship to perceived rate and streaming." J Neurophysiol 97(3): 2230-2238.

Winkler, I. (2007). "Interpreting the mismatch negativity." Journal of Psychophysiology 21: 147-163.

Winkler, I. (2010). In search for auditory object representations. Unconscious memory representations in perception: Processes and mechanisms in the brain. I. Winkle and I. Czigler. Amsterdam, John Benjamins: 71-106.

Winkler, I. and N. Cowan (2005). "From sensory to long-term memory: evidence from auditory memory reactivation studies." Exp Psychol 52(1): 3-20.

Winkler, I. and I. Czigler (2012). "Evidence from auditory and visual event-related potential (ERP) studies of deviance detection (MMN and vMMN) linking predictive coding theories and perceptual object representations." Int J Psychophysiol 83(2): 132-143.

Winkler, I., I. Czigler, E. Sussman, J. Horvath and L. Balazs (2005). "Preattentive binding of auditory and visual stimulus features." J Cogn Neurosci 17(2): 320-339.

Winkler, I., S. Denham, R. Mill, T. M. Bohm and A. Bendixen (2012). "Multistability in auditory stream segregation: a predictive coding view." Philos Trans R Soc Lond B Biol Sci 367(1591): 1001-1012.

Winkler, I., S. L. Denham and I. Nelken (2009). "Modeling the auditory scene: predictive regularity representations and perceptual objects." Trends Cogn Sci 13(12): 532-540.

Winkler, I., E. Sussman, M. Tervaniemi, J. Horváth , W. Ritter and R. Näätänen (2003). "Preattentive auditory context effects." Cogn Affect Behav Neurosci 3(1): 57-77.

Winkler, I., R. Takegata and E. Sussman (2005). "Event-related brain potentials reveal multiple stages in the perceptual organization of sound." Brain Res Cogn Brain Res 25(1): 291-299.

Winkler, I., W. A. Teder-Salejarvi, J. Horváth , R. Näätänen and E. Sussman (2003). "Human auditory cortex tracks task-irrelevant sound sources." Neuroreport 14(16): 2053-2056.

Winkler, I., T. L. van Zuijen, E. Sussman, J. Horvath and R. Naatanen (2006). "Object representation in the human auditory system." Eur J Neurosci 24(2): 625-634.

Winkler, I., T. L. van Zuijen, E. Sussman, J. Horváth and R. Näätänen (2006). "Object representation in the human auditory system." Eur J Neurosci 24(2): 625-634.

Yildiz, I. B. and S. J. Kiebel (2011). "A hierarchical neuronal model for generation and online recognition of birdsongs." PLoS Comput Biol 7(12): e1002303.

Zhuo, G. and X. Yu (2011). Auditory feature binding and its hierarchical computational model. Third international conference on artificial intelligence and computational intelligence Springer-Verlag.

Zwicker, E. and H. Fastl (1999). Psychoacoustics. Facts and Models. Heidelberg, New York, Springer.